Published in final edited form as:

Methods Mol Biol. 2018 ; 1757: 31–47. doi:10.1007/978-1-4939-7737-6_3.

Using the *Candida* Genome Database

Marek S. Skrzypek, Jonathan Binkley, and Gavin Sherlock

Department of Genetics, Stanford University, Stanford, CA 94305-5120

Abstract

Studying *Candida* biology requires access to genomic sequence data in conjunction with experimental information that together provide functional context to genes and proteins, and aid in interpreting newly generated experimental data. The *Candida* Genome Database (CGD) curates the *Candida* literature, and integrates functional information about *Candida* genes and their products with a set of analysis tools that facilitate searching for sets of genes and exploring their biological roles. This chapter describes how the various types of information available at CGD can be searched, retrieved, and analyzed. Starting with the guided tour of the CGD Home page and Locus Summary page, this unit shows how to navigate the various assemblies of the *C. albicans* genome, how to use Gene Ontology tools to make sense of large-scale data, and how to access the microarray data archived at CGD, as well as visualize high-throughput sequencing data through the use of JBrowse.

Keywords

Candida; genome database; expression analysis; gene ontology; GO slim; JBrowse

1. Introduction

The *Candida* Genome Database (CGD; <http://www.candidagenome.org>) was started in 2004 to serve as an online compendium of genomic, genetic and molecular biology information about *Candida albicans*. The approach for CGD itself was derived from the *Saccharomyces* Genome Database (SGD), both in terms of the code base itself, which was modified from SGD's code, but also in terms of the staff who have worked at CGD over the last 13 years – many of them have also worked at SGD at various times. Thus CGD was able to provide a database with a similar look and feel to a community that was already used to SGD, and a similar outlook on the purpose of the database – that is, to serve the experimental scientists working at the bench, to better enable them to understand and interpret their own data, and to design experiments based on the collective knowledge of the community. At the outset, CGD was based on the then newly assembled genomic sequence of strain SC5314 [1], a clinical strain isolated from a patient with disseminated candidiasis [2] (originally from the Squibb Collection, and first referenced in the literature in 1968 [3–5]). The primary goal of CGD was to couple the sequence data with the literature-derived experimental data in a single, easy to navigate web-based resource. Since then, the expanding scope of *Candida* research, facilitated by the progress in sequencing of genomes from other strains and species, has prompted the incorporation of similar data for *C. glabrata* CBS138, *C. parapsilosis* CDC317, and *C. dublinensis* CD36. CGD has also become an archive that

provides access to genomic sequences for other related strains and species, including *C. albicans* WO-1, *C. guilliermondii* ATCC_6260, *C. lusitaniae* ATCC_42720, *C. orthopsilosis* Co 90–125, *C. tropicalis* MYA-3404, *Debaryomyces hansenii* CBS767, and *Lodderomyces elongisporus* NRLL YB-4239.

At the core of CGD lies human curation, a process that involves manually extracting gene-specific experimental information from the published, peer-reviewed literature and associating those annotations with the relevant genomic features. Genes and their annotations are organized in such a fashion that the information is easily browsable, searchable and retrievable for further analysis and perusal. CGD curators also make sure that every piece of information is traceable to its original source, usually a publication in a scientific journal, thus providing access to all available experimental details and their interpretations. CGD also includes a rigorous analysis of orthology between species [6] and of protein domain structure, which allows consistent predictions of functions for genes that have not been experimentally characterized [7]. Thus, CGD provides a structured, unbiased and continuously updated collection of a large variety of experimental results and computational predictions that has become indispensable for *Candida* researchers

In order to ensure a uniform representation of biological information across different organisms, most genome databases use controlled vocabularies to annotate various attributes of genes and gene products. The most widely used vocabulary for capturing the key aspects of gene product biology is the Gene Ontology (GO; <http://www.geneontology.org/>; [8]). GO is a system of standardized terms with defined relationships that describe a primary activity of the gene product (Molecular Function), a broader cellular role the gene product is involved in (Biological Process), and the predominant localization, such as a protein complex, a subcellular structure, or an organelle (Cellular Component). CGD uses GO as the main vocabulary to annotate genes. Another data type that CGD captures, mutant phenotypes, is curated using Ascomycete Phenotype Ontology (APO), a vocabulary developed at *Saccharomyces* Genome Database (SGD) [9] that we adapted to the specific needs of *Candida* biology.

The information in CGD is organized in a system of interlinked web pages, designed with the goal of making them intuitive, easily navigable and user-friendly. However, the sheer complexity of the data presented in CGD can make it difficult for newcomers to find the information they are looking for. This chapter provides help for navigating the site and highlights some of the more recently added features of CGD. We present an overview of the main entry point, the Home Page, and the central organizing principle of the database, the Locus Summary Page. We show how to navigate between the *C. albicans* genome and the genomes of other *Candida* species. We also explain how to perform some of the most common types of analysis that utilize GO annotations. Finally, we show how to access and browse large-scale datasets collected at CGD, including archived microarray datasets, as well as some more recent high throughput sequencing data.

2. Methods

2.1 Exploring CGD Locus Summary Pages

The Locus Summary Page (LSP) serves as the primary organizing principle for gene-specific data in CGD. There is an LSP for each protein- or RNA-coding gene, as well as for each of the other annotated chromosomal features (see Note 1). The LSP provides an overview of what is known about that particular entity and serves as an access point to more detailed information and analysis tools.

1. To access a Locus Summary Page, first open the CGD home page (<http://www.candidagenome.org>) (Fig. 1). Take a look at the yellow menu bar at the top of the page. This bar is present on most CGD pages and facilitates quick navigation to various data search and analysis tools, genome browsers, and literature tools; it also offers links to bulk download tools as well as a variety of community-related information. Hover the mouse over each item to reveal a menu with available options.
2. To find an LSP for a particular feature of interest, enter your query into the “Search our site” box above the banner (see Note 2). After entering an identifier that unambiguously points to a specific entity (for example, enter orf19.922), clicking on the “Go” button will open the LSP for that entity. If your search produces multiple hits, such as a gene name that is used in several *Candida* species represented in CGD (enter *erg11*, as an example), you will see a “CGD Quick Search Result” page that lists the type and number of hits, and a species they come from. Positive hits are hyperlinked to either their respective LSPs, or to an intermediate list of individual hits. Click on an individual hit to open its LSP.
3. The Locus Summary Page has several tabs, whose type and number may differ for different feature types. The Summary tab is open by default. The other tabs on an LSP for a protein-coding gene are Locus History, Literature, Gene Ontology, Phenotype, Homologs, and Protein. Browse through the tabs to see what types of information they contain.
4. Go back to the Summary tab and look at the Basic Information section at the top of the page. It contains a headline-like Description that summarizes the most significant features of the locus. It also lists all the names associated with the locus including the standard name - typically the genetic name under which the gene was first published. If there are other names by which the gene has been referred to in the literature, they are listed as aliases. The identifier assigned during genome sequencing is listed as Systematic Name along with the name of the reference strain used in the sequencing project (see Note 3). The Basic Information section allows easy navigation to genes in other organisms. Click on any hyperlinked name listed among “Orthologous genes in *Candida* species” to open its LSP in CGD. You can also click on the “View ortholog cluster” link, which will show the orthologs from 15 *Candida* species in their genomic context, a report produced by the Candida Gene Order Browser [6]. To explore orthologs

from other fungal species, select a gene from “Ortholog(s) in non-CGD species” and you will access information at other resources: AspGD, Broad Institute, PomBase, and SGD, for genes from *A. nidulans*, *N. crassa*, *S. pombe*, *S. cerevisiae*, respectively (see Note 4). At the bottom of this section, there is a thumbnail showing the chromosomal location of the gene, hyperlinked to GBrowse [10,11], which provides a graphical interface to inspect the genomic context of the gene.

5. The GO Annotations and Mutant Phenotypes sections show current information about the function of the gene. For more detailed information, including the references on which these annotations are based, open the Gene Ontology or Phenotype tab, respectively. GO annotations on the Summary tab are divided by the GO aspect (Molecular Function, Biological Process, Cellular Component) and by the annotation method. Manually curated annotations are assigned by a CGD curator based on reading of the scientific literature, while computational annotations are produced by transferring experiment-based GO annotations from orthologous genes in other species, or by predictions based on domain structure. Each annotation is accompanied by an evidence code that indicates the reason behind the annotation, for instance, IMP means Inferred from Mutant Phenotype. Click on any evidence code to see a table of all the evidence codes and their definitions. Annotations based on sequence similarity, genetic or physical interactions also list the source gene(s) and the organism. Click on any gene name to see a report for that gene in CGD or in an external database. Each GO term itself is hyperlinked to another CGD page that provides more information, including term definitions, a diagram depicting the relevant segment of the ontology, and a list of other genes in CGD that are also annotated with that term. Similarly, each mutant phenotype is hyperlinked to a page that lists all other genes in CGD that display the same phenotype.
6. The Sequence Information section shows the basic data about the gene (chromosomal coordinates, intron-exon structure), but also provides easy access to sequences and sequence analysis tools. Open the drop-down menu next to Retrieve Sequences to see available options that include retrieval of DNA sequence in several configurations and, for protein-coding genes, for the predicted protein sequence as well. Similarly, open the Sequence Analysis Tools drop-down menu to start BLAST searches, restriction analysis or primer design tools. More analysis tools are available from the banner on top of the page; click on Search, Sequence, Tools, or Download to see the options. In addition, at the bottom of the Sequence Information section there are links to sequence data available from external resources, such as GenBank, UniProt and others.

2.2 Explore *Candida* genomes using genome browsers

Genome browsers allow visualization of massive amounts of information in an easily understandable graphical format. The genomic sequence is overlaid with functional annotations, thus providing an overview of entire regions and, at the same time, a quick access to individual genome features. The genome browsers implemented in CGD are

Generic Genome Browser, or GBrowse [10,11], and JavaScript-based Genome Browser JBrowse [12].

2.2.1 Browsing the chromosomal features using GBrowse

1. Access GBrowse from the menu bar at the top of most CGD pages. When you hover the cursor over the GBrowse option, a menu unfolds that lets you select one of the *Candida* species available in CGD (see Note 5).
2. In the Search panel of the main GBrowse window (Fig. 2), you can search for a particular gene of interest by entering its name in the Landmark or Region box. The search tool also works for other types of queries (see Note 6).
3. You can switch to a different Data Source by selecting one of the historic *C. albicans* assemblies, one of the other *Candida* species available in CGD, or you can change the browser to display protein data.
4. The Search panel also provides the ability to scroll up and down along the chromosome using the yellow <<, <, >, >> buttons, or zoom in and out using the yellow – or + buttons. You can also select a desired zoom level from a pull-down menu. In addition, the Search panel provides tools to generate a restriction map of the displayed region, download the sequence and save the current display as a snapshot
5. The Overview and Region panels show schematic diagrams of the entire chromosome, or the selected region, with the highlight representing the currently displayed region. The diagrams can be used for navigation: a single click re-centers the display on the clicked location; a click-and-drag specifies a new region to display.
6. The Details panel with the default settings displays all annotated sequence features, including protein coding genes, tRNA and snRNA genes, transposons, and centromeres. The features are shown as horizontal bars with pointed ends and their color indicates their orientation, where applicable: features encoded on the Watson strand are red and those on the Crick strand are blue. Depending on the zoom level, each feature may be labeled with its name and description. Hovering over a feature generates a window with more information and clicking on a feature opens its Locus Summary page.
7. The content displayed in the Details panel is highly customizable. Click on Select Tracks tab on top, or Select Tracks button at the bottom to find the list of available datasets that can be overlaid on the sequence. Check or uncheck the desired boxes and hit Back to Browser to see the results (see Note 7)

2.2.2 Browse and analyze large-scale datasets using JBrowse

1. Launch JBrowse from any CGD feature's Locus Summary Page by clicking on the JBrowse logo about halfway down the page. The JBrowse window will be centered upon that feature (Fig. 3). Alternately, mouse-over “JBrowse” on the top menu bar of any CGD page, then select and click your organism of interest in the

drop-down menu. The browser window will be centered on an arbitrary location of the genome (see Note 8).

2. The top menu bar of the browser has 4 items:
 1. JBrowse: Displays version information with links to the Generic Model Organism Database (GMOD) websites.
 2. File: Load data and customize tracks, as described below.
 3. View: Highlight features of interest, or resize quantitative data tracks.
 4. Help: Display version information and general browser help.
3. Beneath the top menu bar are the navigation controls. The arrow and magnification buttons allow panning and zooming of the display. The chromosome pull-down menu to the right of the buttons changes the chromosome displayed. The next box to the right of that displays the current location, but it can also be used to search by location or feature name simply by clicking and entering text (best search results will be obtained by entering the systematic name of the target feature). Just above the buttons and search box is a linear representation of the current chromosome, with a draggable red box over the current location. Change the current location by clicking and dragging the red box. Change the size of the region displayed by clicking on either side of the box and resizing it.
4. The main browser window displays data tracks of various types. A label in the far left of each track shows the track name, and has a pull-down menu with technical information about the track. You can change the relative position of the track in the window by clicking on the label and dragging it up or down to the desired position. The main types of tracks are:
 1. Sequence Tracks display nucleotide sequence of both strands, as well as 6-frame translated amino acid sequences. Note that the window may need to be zoomed-in considerably for sequence tracks to be displayed.
 2. Feature Tracks show all transcribed features annotated at CGD. Features encoded on the “W” (plus) strand are displayed in red, and features on the “C” (minus) strand in blue. Clicking on a feature brings up an information window, and clicking on the feature name within that window opens the CGD Locus Summary Page for that feature.
 3. Quantitative Tracks graphically display various types of numeric data – for example, expression levels, degree of conservation, or density of aligned high-throughput sequence reads. Most quantitative data are shown on a log scale, but this can be changed in the information pull-down menu at the far left of the track.
 4. Alignment Tracks show the genomic alignments of sequence reads from a variety of high-throughput experiment types: DNA-Seq, RNA-Seq, ChIP-Seq, etc. (see Note 9).

5. Variation Tracks display small variations (SNPs or indels) between two different genomes (or between the two haplotypes of a diploid genome, such as that of *C. albicans* SC5314).
5. To choose which tracks to display, click on the “Select Tracks” tab on the upper-left side of the main JBrowse window. This brings up the Track Selection Menu (Fig. 4). The many available tracks are listed in the right panel of the menu. To select tracks, click the checkbox next to the item description. The left panel of the selection menu gives a number of filtering criteria to help narrow down the track choices. For example, you might filter by experimental technique (RNA-Seq, DNA-Seq, etc.), experimental condition (oxidative stress, pH, control, etc.), or source publication (either by first author or PubMed ID). Select “Currently Active” under “My Tracks” to show only the tracks that are currently selected and displayed in the main window. The “Clear All Filters” button above the right panel removes any filters applied from the left panel (but does not deselect tracks), returning the display to all available tracks. You can also find tracks using the search box above the right panel: for example, searching *C. albicans* SC5314 tracks with the text “stress” shows selections from both oxidative stress and nitrosative stress experiments. Clicking the “Back to Browser” button above the right panel executes any track changes you made and returns to the main JBrowse window.
6. You can display all instances of a particular sequence or motif as a separate track. From the File menu in the top menu bar, select “Add sequence search track”, and enter the query sequence (either as plain text or a regular expression). You can search using either nucleotide or amino acid sequence, and you can search on either or both strands.
7. You can combine information from separate tracks into a custom “combination” track. For example, you may want to create a Search Track for a given sequence motif, as above, but only display the instances that occur in genes included in the “Transcribed Features” track. From the File menu in the top menu bar, select “Add combination track”. A new “empty” track is created. Click on the information labels of the tracks you wish to combine, drag them over to the information label of the new combination track, and release to combine (the label will appear red while you are dragging it, and turn green when it is in position over the combination track). After the different tracks have been added, a menu pops up to select details of the combination (intersection, union, etc.). For the above example, you would select “Intersection”.
8. You can load your own data for display as tracks in CGD JBrowse. Accepted data formats include BAM, GFF3, Wiggle/BigWig, or VCF. Data can be either in files on your local system or from a remote web server. From the File menu in the top menu bar, select “Open track file or URL”, and add local files to the left box, or URLs to files from web servers to the right box. Note that in order to display correctly, the data must be annotated with the sequence coordinates of the current CGD genome version.

2.3 GO Term Finder and GO Slim Mapper, tools for GO Analysis

The GO vocabularies are structured as hierarchies, with more general terms, so-called “parents”, encompassing more specific “child” terms. It is a standard requirement for the curators to assign the most specific (granular) term that the evidence presented in the publication supports. Within the hierarchical structure of the ontology, annotation to a child term implies that annotation to its parent term is also correct. This feature of GO has to be taken into account during analysis of large-scale data, when it is necessary to identify common biological features in a set of genes that are, for instance, co-regulated in a microarray experiment. Finding statistically significant similarities in GO annotations for multiple genes requires a tool, GO Term Finder, that is able to move up the GO hierarchy from the specific terms used to annotate the genes in a list to their GO parent terms that the genes may have in common. Given the complexity of GO, it is often desirable to group genes into broader categories, using only high-level terms. The GO Slims are such lists of high-level terms from each ontology branch (Molecular Function, Biological Process, and Cellular Component), carefully selected to cover most of the curated GO information. The tool, GO Slim Mapper, allows users to pick such GO Slim terms and then map a set of genes to them.

2.3.1 Using GO Term Finder

1. Open the GO Term Finder page by selecting it from the options in the Search or GO pull-down menus in the menu bar (see Note 10).
2. Select the species in Step 1; the default species is *Candida albicans*, but you can run this analysis for other CGD-curated species: *C. glabrata*, *C. parapsilosis* or *C. dublinensis*.
3. In Step 2, enter a list of gene names. You can either type the name of the genes in the input box or upload a file that contains the list. Either genetic names (CGD Standard Names, e.g., 'AAFT') or systematic names ('C3_06470W_A', or orf19 identifiers, e.g., 'orf19.7436') may be used (see Note 11). As an example, use these genes: MOB2 GAL10 AGE3 ECM33 MNS1 SOG2 VRG4 DPM1 SAC1 CBK1.
4. In Step 3, select one of the three branches of GO (biological process, molecular function, or cellular component) by checking the boxes. The tool only searches one of the three branches at a time.
5. Click the Search button after Step 3 to use the default settings or go further down to Steps 4 and 5 to specify and customize your background set and/or refine the types of annotations in your background set.
6. You may change your background set in Step 4. The default background set includes all the genes in the database that have at least one GO annotation. You can also customize the background set by choosing which feature type(s) it should include.

7. In Step 5 you can deselect specific types of GO annotations that should not be used for calculations. By default, annotations collected by all methods and with all types of evidence are included,
8. The results page displays the significant shared GO terms (or their parents) in both graphic and table form, within the set of genes entered on the previous page.
9. The graphic shows the GO tree that includes terms used directly or indirectly in annotations for the genes in your list. The terms are color-coded to indicate their statistical significance (p-value score). Genes associated with the GO terms are shown in gray boxes, with links to their respective Locus Summary pages.
10. The table below the graph lists each significant GO term, the number of times the GO term is used to annotate genes in the list and the number of times that the term is used to annotate genes in the background set (see Note 12).
11. Additional columns list the p-value, the False Discovery Rate (FDR), and a list of all the genes annotated, either directly or indirectly, to the term. FDR is an estimate of the percent chance that a particular GO term might actually be a false positive. It represents the fraction of the nodes with p-values as good or better than the node with this FDR that would be expected to be false positives.
12. The statistical significance of the association of a particular GO term with a group of genes in the list is indicated by the p-value: the probability of seeing at least x number of genes out of the total n genes in the list annotated to a particular GO term, given the proportion of genes in the whole genome that are annotated to that GO Term. The closer the p-value is to zero, the more significant the particular GO term association with the group of genes is (i.e. the less likely to occur by chance).

2.3.2 Using GO Slim Mapper

1. Select GO Slim Mapper from the options in the Search or GO pull-down menus in the menu bar. In Step1 select the species for your analysis (default is *Candida albicans*). In Step 2 you can type or paste your list of genes or upload them as a file. As an example, you can use the same gene list as previously: MOB2 GAL10 AGE3 ECM33 MNS1 SOG2 VRG4 DPM1 SAC1 CBK1.
2. In Step 3, use the pull-down menu to select the GO Set Name: GO Slim Component, GO Slim Function, or GO Slim Process. The list of terms from the selected set appears in the window in Step 4.
3. In Step 4 you can specify which particular term, or terms, you want to use. The default setting is “Select ALL Terms,” but you can highlight only terms you are interested in.
4. In optional Step 5 you can exclude certain types of annotations, e.g., computational and high-throughput experiments.
5. Click Search to start the process; note that long lists of genes will take significant time to analyze.

6. Results appear in a table with three columns: the GO Slim terms chosen, with a link to graphical depiction of that branch of GO, the percentage of genes in your list annotated to each term, and the genes from your list that are annotated to that term. You can also download the results in a tab-delimited file.

2.4 Browsing Large-Scale Data and GeneXplorer

CGD has been collecting large-scale datasets from its inception. The datasets from authors' websites or from supplementary materials that accompany publications are archived in the Datasets page accessible from the Download menu in the banner, thus offering a convenient one-stop shop for datasets pertaining to *Candida* that can be then imported into one's own tools for analysis. The microarray results have been built into GeneXplorer, a web tool that allows browsing and analysis of the expression data at the CGD site, with the benefit of an instant access to functional annotations [13].

1. To access the archived datasets, click on the Download pull-down menu in the top banner and select Datasets. The resulting page lists all the publications from which the datasets are archived in CGD, ordered by year and author (Fig. 5). All the references have icons that lead to the abstract, full-text articles and to the original data website, as well as to the download directory at CGD. The microarray-containing references have a pair (or pairs) of green-blue icons. Click the one on the right that is labeled “.pcl” and you will get the entire datasets in the pcl format, a standard format that makes the dataset amenable to analysis by many popular software suites. Find a publication of interest and click the icon left to the “.pcl” icon to see the data in GeneXplorer (Fig. 5).
2. The GeneXplorer page (Fig. 6) shows a heat map image of the clustered expression profiles from the selected publication, where increased or decreased gene expression is shown in shades of red or green, respectively. Zoom in and out using the + or – buttons above the cluster. Select any particular region by dragging over it. Click on a profile for a particular gene to see the expression patterns of the most similarly or dissimilarly expressed genes. Find a gene of interest by entering its name into the “Search for” box. Click on any gene name to go to its Locus Summary page.

3. Notes

1. Several entities other than protein-coding genes have their Locus Summary pages in CGD, including various RNA genes, such as tRNA, rRNA, snRNA, as well as centromeres, telomeres, repeated sequences, and other annotated chromosomal features.
2. The query entered into the search box may be a *Candida* gene or protein name (standard or systematic name, or an alias), author or colleague name, PubMed ID, or any keyword (such as a functional term or phenotype). It can even be a name of an ortholog from one of the non-CGD species. When there are multiple hits, a list of matches is displayed. The “Search our site” box is case-insensitive and accepts a wildcard character *. For example, enter “act*” to retrieve any

piece of data starting with “act”. Also, the search box has an autocomplete feature, which provides suggestions when you start typing your query.

3. For *C. albicans*, the systematic name shown on the Locus Summary page is always the name of the haplotype A allele, as denoted by an “A” suffix, with the corresponding haplotype B allele listed below. This new systematic name is based on the known chromosomal location and haplotype, and it consists of the chromosome (C1-C7 and CR for the eight nuclear chromosomes, CM for the mitochondrial chromosome), a unique number indicating the order of features along chromosomes, the strand (W for Watson or C for Crick) and the haplotype (A or B) [14]. Since the systematic names from previous assemblies of *C. albicans* genome (so called “orf19” names) continue to be widely used, the Assembly 19/21 identifier is also listed. The systematic Identifiers from earlier assemblies are included among the aliases.
4. The ortholog mappings among *Candida* strains, and between *Candida* strains and *S. cerevisiae*, are derived from the curated syntenic groupings at the Candida Gene Order Browser (CJOB) [15]. The ortholog mappings between *Candida* strains and *S. pombe*, *A. nidulans*, and *N. crassa* are made by pairwise comparisons using the InParanoid software [16].
5. Gbrowse is also accessible from any Locus Summary page. Click on a thumbnail in the Chromosomal Location section and a browser window will open, centered on the gene of the page and 10 kbp upstream and downstream. Note that in *C. albicans* there are two alleles and the two thumbnails lead to either haplotype A or haplotype B chromosomes.
6. The GBrowse search tool recognizes several types of landmarks, such as a chromosome (“Ca22chr1A_C_albicans_SC5314”) or coordinates (“Ca22chr1A_C_albicans_SC5314:724532..795083”), a gene name (“*ATP1*”), a systematic ORF name (“C1_04610W_A”), an alias (“orf19.6854”). You can also use keywords (“ATP synthase”) and wildcards (“ATP*”). When the search results include multiple hits, the browser will show a diagram of all chromosomes with landmarks matching your query and a table with a list of hits. Click on one of the landmarks, either in the diagram or in the table, to see that region to the Details panel of the browser.
7. There is a set of icons to the left of the track name that provide additional controls. Hovering over each icon produces a pop-up window that explains its function.
8. CGD has aligned high-throughput sequence data to each of the two haplotypes of *C. albicans* SC5314 independently, and the two haplotypes are displayed in JBrowse separately. This is primarily to reduce alignment errors for short-read sequences, but it also allows sequence variation to be viewed in the context of either haplotype. By default, JBrowse opens with the “A” haplotype displayed.
9. The browser window may need to be zoomed in to view alignments, and scrolling may be necessary to view all the aligned reads. Browser speed and

performance may be greatly diminished if more than one alignment track is displayed. For each alignment, CGD also provides two quantitative tracks, summarizing read coverage and read density. These may be preferred if more than one dataset are to be viewed or compared.

10. Links to both GO Term Finder and GO Slim Mapper appear at the bottom of many pages with results of searches that produce a list of genes. Selecting those links open the respective tool with the gene list already pre-loaded into the input box.
11. When a name is entered that is an alias for one gene or feature, the program will map the name to that gene. If the name is an alias for more than one gene but not a standard or systematic name for any genes, the program will present a list of possible mappings. The user can decide which gene was intended and edit the input.
12. Because the frequency of any given annotation within the background set is compared against the frequency of the annotation within the query set (input), the choice of background set affects the significance of the results that are returned by the tool. Please note that the specific background set of genes that was used in the absence of any user-defined set (the default background set) has changed over time.

Acknowledgements

This work was supported by National Institute of Dental and Craniofacial Research at the US National Institutes of Health (grant no. R01 DE015873).

5. References

1. Jones T, Federspiel NA, Chibana H, et al. The diploid genome sequence of *Candida albicans* Proc Natl Acad Sci USA. 2004; 101(19):7329. 10.1073/pnas.0401648101. [PubMed: 15123810]
2. Gillum AM, Tsay EY, Kirsch DR. Isolation of the *Candida albicans* gene for orotidine-5'-phosphate decarboxylase by complementation of *S. cerevisiae* *ura3* and *E. coli* *pyrF* mutations Mol Gen Genet. 1984; 198(2):179. [PubMed: 6394964]
3. Aszalos A, Robison RS, Lemanski P, et al. Trienine, an antitumor triene antibiotic J Antibiot (Tokyo). 1968; 21(10):611. [PubMed: 5715671]
4. Maestrone G, Semar R. Establishment and treatment of cutaneous *Candida albicans* infection in the rabbit Naturwissenschaften. 1968; 55(2):87. [PubMed: 5703113]
5. Meyers E, Miraglia GJ, Smith DA, et al. Biological characterization of prasinomycin, a phosphorus-containing antibiotic Appl Microbiol. 1968; 16(4):603. [PubMed: 5647519]
6. Fitzpatrick DA, O'Gaora P, Byrne KP, et al. Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser BMC Genomics. 2010; 11:290. 10.1186/1471-2164-11-290. [PubMed: 20459735]
7. Inglis DO, Arnaud MB, Binkley J, et al. The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata* Nucleic Acids Res. 2012; 40:D667. 10.1093/nar/gkr945. [PubMed: 22064862]
8. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Nat Genet. 2000; 25(1):25. 10.1038/75556. [PubMed: 10802651]

9. Engel SR, Balakrishnan R, Binkley G, et al. Saccharomyces Genome Database provides mutant phenotype data Nucleic Acids Res. 2010; 38(Database issue):D433. 10.1093/nar/gkp917. [PubMed: 19906697]
10. Stein LD. Using GBrowse 2.0 to visualize and share next-generation sequence data Brief Bioinform. 2013; 14(2):162. 10.1093/bib/bbt001. [PubMed: 23376193]
11. Stein LD, Mungall C, Shu S, et al. The generic genome browser: a building block for a model organism system database Genome Res. 2002; 12(10):1599. 10.1101/gr.403602. [PubMed: 12368253]
12. Buels R, Yao E, Diesh CM, et al. JBrowse: a dynamic web platform for genome visualization and analysis Genome Biol. 2016; 17:66. 10.1186/s13059-016-0924-1. [PubMed: 27072794]
13. Rees CA, Demeter J, Matese JC, et al. GeneXplorer: an interactive web application for microarray data visualization and analysis BMC Bioinformatics. 2004; 5:141. 10.1186/1471-2105-5-141. [PubMed: 15458579]
14. Skrzypek MS, Binkley J, Binkley G, et al. The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data Nucleic Acids Res. 2017; 45:D592. 10.1093/nar/gkw924. [PubMed: 27738138]
15. Maguire SL, OhEigeartaigh SS, Byrne KP, et al. Comparative genome analysis and gene finding in *Candida* species using CGOB Mol Biol Evol. 2013; 30(6):1281. 10.1093/molbev/mst042. [PubMed: 23486613]
16. Ostlund G, Schmitt T, Forslund K, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis Nucleic Acids Res. 2010; 38:D196. 10.1093/nar/gkp931. [PubMed: 19892828]

Fig. 1. Locus Summary Page for *C. albicans* ERG2 - Basic Information section; red arrows point to the search box present on most CGD pages, and the tabs, that allow switching between different types of information.

Fig. 2.
 GBrowse main window, showing genomic features for a region of chromosome 5.

Fig. 3. JBrowse main window, showing short reads from an RNA-Seq experiment aligned against a region of the genome. Read color (red or blue) allows inference of the expressed strand, and corresponds to the direction of known genes.

Fig. 4. JBrowse track selection window. Different datasets can be selected for display.

- 2013:
- Aoki W, et al. (2013) Elucidation of potentially virulent factors of *Candida albicans* during serum adaptation by using quantitative time-course proteomics. *J Proteomics* 91:417-29

 - Carlisle PL and Kadosh D (2013) A genome-wide transcriptional analysis of morphology determination in *Candida albicans*. *Mol Biol Cell* 24(3):246-60

 - Chakraborty U, et al. (2013) A stable hybrid containing haploid genomes of two obligate diploid *Candida* species. *Eukaryot Cell* 12(8):1061-71

 - Chen YY, et al. (2013) Dynamic transcript profiling of *Candida albicans* infection in zebrafish: a pathogen-host interaction study. *PLoS One* 8(9):e72483

 - Cheng S, et al. (2013) Profiling of *Candida albicans* gene expression during intra-abdominal candidiasis identifies biologic processes involved in pathogenesis. *J Infect Dis* 208(9):1529-37

Fig. 5.
Example entries in the Dataset Archive page. The red arrow points to the icon leading to the GeneXplorer window.